Self-Training with Plant Village

Piyush S. Waradkar
CS 584
Illinois Institute of Technology
Chicago, IL 60616
pwaradkar@hawk.iit.edu

Abstract

This project explores the application of a Self-Trained Convolutional Neural Network (CNN) for plant disease in the field of agricultural technology, specifically for the task of plant disease classification. Utilizing the Plant Village dataset, which comprises 20,000 labelled images of various plant diseases and healthy leaves, the project aims to enhance the accuracy and efficiency of disease diagnosis through a semi-supervised learning approach.

Central to this endeavour is the development of a CNN model, built and trained from scratch, tailored to the specific nuances of the Plant Village dataset. The initial phase involves training the CNN on a smaller subset of labelled images, employing this model to generate pseudo-labels for a larger set of unlabelled images. These pseudo-labels, particularly those predicted with high confidence, are then assimilated back into the training process, in an iterative self-training cycle. This methodology not only utilizes the available labelled data but also capitalizes on the untapped information present in the unlabelled data, potentially leading to a more refined and robust classification model.

The project's objective is to demonstrate how a CNN, developed without the foundational aid of transfer learning but enhanced through self-training, can effectively improve plant disease classification. This approach is particularly significant given the limited amount of labelled data typically available in specialized domains like agricultural disease diagnosis. The successful implementation of this method holds promise for real-world applications, where it can aid farmers and agronomists in timely and accurate disease detection, thereby contributing to improved crop yields and sustainable farming practices.

.

Background Review

1. Importance of Agricultural Disease Diagnosis:

Agricultural disease diagnosis is crucial for maintaining high crop yields and ensuring food security. The early and accurate identification of plant diseases allows for timely intervention, preventing widespread crop damage.

The global significance of this issue is underscored by its direct impact on food supply chains, farmer livelihoods, and the overall health of agricultural ecosystems.

2. Advancements in Machine Learning and Computer Vision:

Recent advancements in machine learning, particularly in computer vision, have revolutionized many fields, including agriculture.

Computer vision techniques, such as image classification and pattern recognition, have been increasingly applied to agricultural contexts, enabling automated and precise disease diagnosis in plants.

3. Plant Disease Classification Challenges:

One of the significant challenges in automated plant disease classification is the variability and complexity of disease symptoms, which can be subtle and easily confused with other stress factors like nutrient deficiencies.

Another challenge is the limited availability of high-quality, labelled datasets specific to various plant diseases, which are essential for training accurate machine learning models.

4. The Plant Village Dataset:

The Plant Village dataset, hosted on Kaggle, is a comprehensive collection of labelled images representing a wide range of plant diseases and healthy plant leaves. It serves as a crucial resource for training and testing plant disease classification models.

The dataset includes over 20,000 labelled images, making it one of the largest and most diverse of its kind, offering an invaluable tool for researchers and practitioners in the field of agricultural technology.

5. Semi-Supervised Learning and Self-Training:

Semi-supervised learning, a machine learning approach that utilizes both labelled and unlabelled data, is particularly relevant in scenarios where labelled data is scarce or expensive to obtain.

Self-training, a form of semi-supervised learning, involves using a model trained on labelled data to generate pseudo-labels for unlabelled data. These pseudo-labels are then used to further train the model, potentially enhancing its performance and generalization capabilities.

6. Relevance to the Project:

Given the limitations in labelled data and the complexity of plant diseases, the semi-supervised approach using self-training presents a promising solution.

This project, by applying a self-trained CNN to the Plant Village dataset, aims to address the challenges in plant disease classification and demonstrate the effectiveness of this approach in agricultural technology.

Problem definition

1. The Challenge of Accurate Plant Disease Classification:

The primary problem your project addresses is the accurate classification of plant diseases using machine learning techniques. Accurate disease classification is vital for effective crop management and disease control, directly influencing crop yield and food security.

Plant diseases can present a variety of symptoms that are often subtle and can be easily confused with other factors, such as environmental stress or nutrient deficiencies. This variability makes accurate classification challenging.

2. Limitations in Existing Approaches:

Many existing approaches rely heavily on manual diagnosis by experts, which can be time-consuming, costly, and subject to human error.

Automated methods that do exist often depend on large volumes of high-quality, labelled data. However, acquiring such datasets is resource-intensive and not always feasible, particularly for less common plant diseases.

3. The Need for Improved Techniques in Resource-Constrained Scenarios:

There is a significant need for methods that can perform well even with limited labelled data. This is particularly relevant in resource-constrained scenarios, such as small-scale farming or in developing countries, where access to large labelled datasets and expert knowledge might be limited.

4. Gap in Leveraging Unlabelled Data:

A substantial amount of agricultural data exists in unlabelled forms. However, many machine learning models fail to utilize this abundant resource effectively, leading to a gap in maximizing the potential of available data.

5. Specific Problem Addressed by This Project:

This project aims to tackle the issue of plant disease classification by employing a semi-supervised learning approach, specifically using a self-training methodology with a CNN. The focus is on leveraging both labelled and unlabelled data from the Plant Village dataset to enhance the model's ability to accurately classify plant diseases.

The project addresses the challenge of training a robust and accurate classification model with a limited amount of labelled data and explores the effectiveness of pseudo-labelling in improving the model's performance.

Algorithms or Method used to solve the problem

The project utilizes a custom Convolutional Neural Network (CNN), PlantCNN, and adopts three distinct approaches to solve the problem of plant disease classification using the Plant Village dataset

Convolutional Neural Network (CNN) - PlantCNN:

1. Architecture Overview:

i. Convolutional Layers: The network consists of four convolutional layers. Each convolutional layer is followed by batch normalization and a ReLU activation function, enabling the network to learn complex features from the input images.

ii. Pooling Layers: Following each convolutional layer, there is a max pooling layer designed to reduce the spatial dimensions of the output, thereby reducing the computational load and extracting dominant features.

iii. Global Average Pooling: A global average pooling layer follows the convolutional layers, reducing each feature map to a single value, which helps to decrease the number of parameters and reduce overfitting.

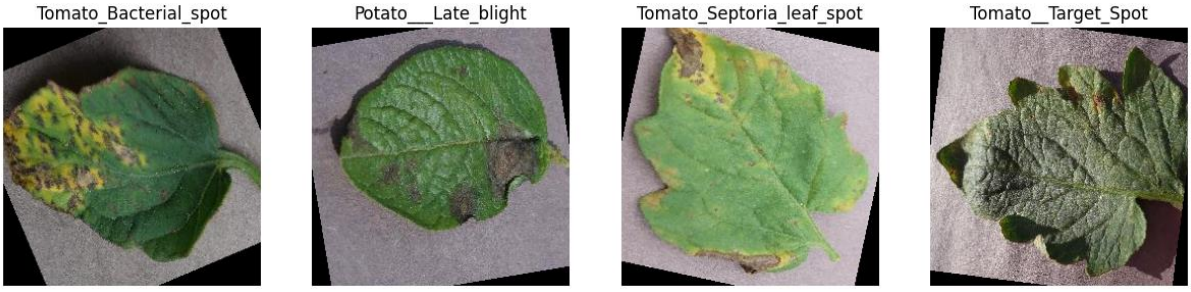
iv. Fully Connected Layer: A fully connected layer at the end of the network maps the learned features to the final classification output, corresponding to the different plant diseases.

v. Dropout Layer: A dropout layer is included to prevent overfitting by randomly dropping units from the neural network during training.

2. Data Preparation and Training:

The Plant Village dataset is pre-processed, which includes steps like normalization and resizing of images, to prepare them for efficient training with the CNN.

Train Images:



Validate Images:

Algorithmic Approaches

1. Direct Training on Labelled Data:

In this supervised learning approach, PlantCNN is trained solely on the labelled portion of the Plant Village dataset. This establishes a baseline for the model's performance, which is crucial for assessing the benefits of the subsequent semi-supervised approaches.

2. Single Round of Self-Training:

After initial training, PlantCNN is employed to predict labels for the unlabelled dataset. High-confidence predictions are used as pseudo-labels and added to the original training dataset, thus expanding it.

The model is then retrained on this new, expanded dataset. This semi-supervised learning approach aims to enhance the model's performance by leveraging additional data.

3. Iterative Self-Training:

Building upon the second approach, this method involves multiple cycles of self-training. After each cycle, the model is used to generate new pseudo-labels from the remaining unlabelled data, which are then used to further expand the training dataset.

This iterative process continues for a predetermined number of cycles or until no significant improvement is observed. The objective is to assess whether continuous self-training can yield progressively better results compared to the single round of self-training.

Each of these approaches serves a specific purpose: establishing a performance baseline, gauging the impact of expanding the training set with pseudo-labels, and exploring the benefits of iterative learning. The comparative analysis of these methods will provide valuable insights into the effectiveness of semi-supervised learning in the context of plant disease classification using CNNs.

Instructions on the Experiment:

Data Properties and Preparation:

Dataset: The Plant Village dataset, consisting of 20,000 images, is used. This dataset includes various plant diseases and healthy leaf images.

Data Division: The dataset is divided into a labelled set for initial training and an unlabelled set for the self-training phase. A validation set is also separated from the labelled set.

Image Processing: Images undergo preprocessing which includes normalization and resizing. Data augmentation techniques such as random rotations and horizontal flips are applied to the training set to improve model robustness.

Hyperparameter Settings:

Batch Size: 32. This determines the number of samples processed before the model is updated.

Learning Rate: Initially set to 0.001. The learning rate is a critical parameter in the training of neural networks, influencing the convergence speed and quality.

Optimizer: Adam, known for its effectiveness in handling sparse gradients and adaptive learning rates.

Loss Function: CrossEntropyLoss, suitable for multi-class classification tasks.

Dropout Rate: 0.5, used in the fully connected layer to prevent overfitting.

Training and Testing Process:

Initial Model Training: The PlantCNN model is first trained on the labelled data subset. This phase establishes a baseline performance for the model.

Self-Training Approaches:

Single Round of Self-Training: After initial training, the model predicts labels for the unlabelled data. High-confidence predictions are used as pseudo-labels and added to the training set. The model is then retrained on this expanded dataset.

Iterative Self-Training: In this approach, the self-training process is repeated multiple times. After each round, the model is retrained with an incrementally expanded training set, integrating new pseudo-labels.

Validation: The model's performance is continually assessed on the validation set to monitor improvements and guide adjustments.

Testing: Post the final training iteration, the model is evaluated on the validation set to assess its performance, particularly in comparison to the initial baseline.

Hardware and Software Environment:

The experiments are conducted using Python and PyTorch, a popular deep learning library. A CUDA-capable GPU is utilized for training.

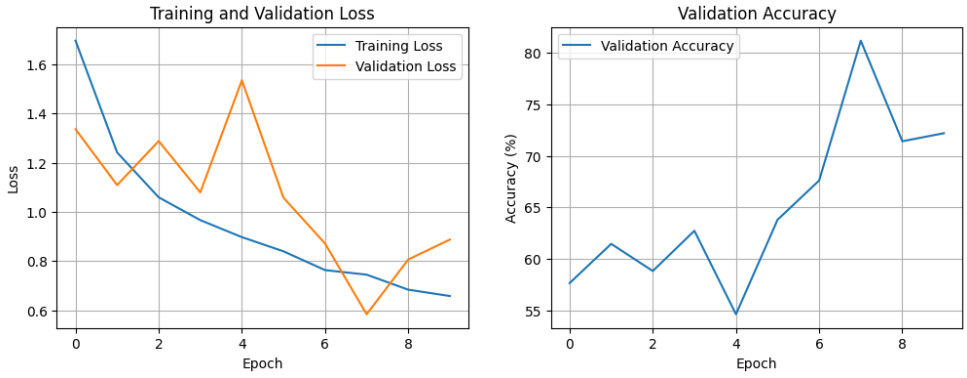
Result Analysis.

Combined Analysis of Methods 1, 2, and 3:

The experimental evaluation of three distinct training methods on the Plant Village dataset reveals insightful trends in model performance and generalization capabilities.

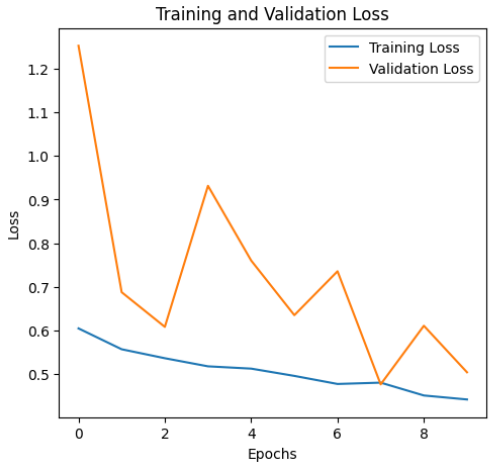
Method 1 (Direct Training on Labelled Data):

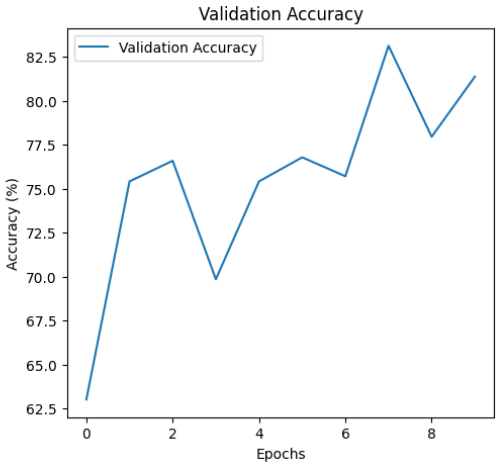
- Established a baseline with a training loss decreasing to 0.6582 and a validation loss fluctuating, indicating potential overfitting.
- Achieved a peak validation accuracy of 72.20% and a test accuracy of 69.10%.
- Precision and F1 score were moderate at 0.7594 and 0.5842, respectively, suggesting room for improvement in model generalization.



Method 2 (Single Round of Self-Training):

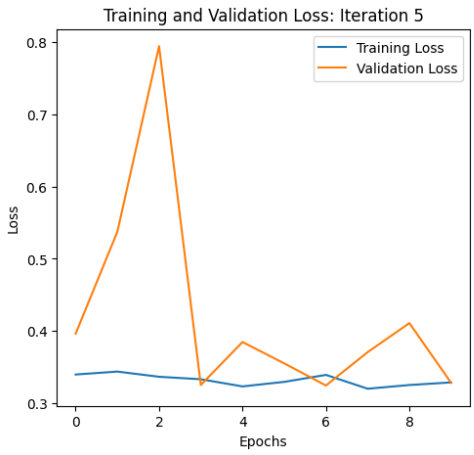
- Improved upon the baseline, with a reduction in validation loss to 0.5037 and a more stable loss trend.
- Validation accuracy peaked at 83.12%, with test accuracy significantly higher at 80.86% compared to method 1.
- Notable increases in precision and F1 score to 0.8118 and 0.7592, respectively, indicating enhanced precision and recall balance.

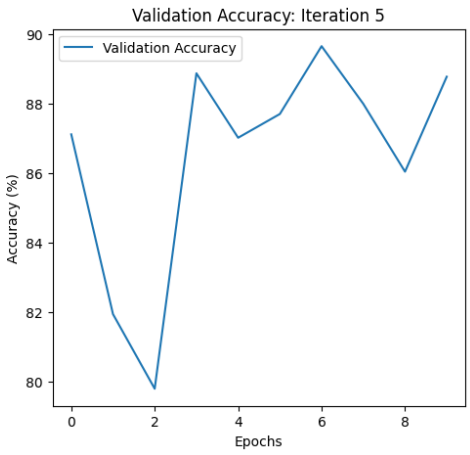




Method 3 (Iterative Self-Training):

- Demonstrated the best performance with a consistent decrease in training loss, the lowest validation loss of 0.3243, and the highest validation accuracy of 89.66% in the final iteration.
- Test accuracy rose to 86.64%, surpassing both method 1 and 2, substantiating the effectiveness of iterative self-training.
- Precision and F1 score reached the highest values at 0.8833 and 0.8103, respectively, underscoring the robustness and reliability of the model.





Approach	*Test Loss*	*Test Accuracy*	*Precision*	*F1 Score*
Approach 1	0.9953	69.10%	0.7594	0.5842
Approach 2	0.5810	80.86%	0.8118	0.7592
Approach 3	0.4071	86.64%	0.8833	0.8103

References

[1] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, Yury Maximov (2022) Self-Training: A Survey. https://arxiv.org/abs/2202.12040

[2] Dough Steen (2020). A Gentle Introduction to Self-Training and Semi-Supervised Learning. https://towardsdatascience.com/a-gentle-introduction-to-self-training-and-semi-supervised-learning-ceee

[3] Mikey Taylor (2020) Computer Vision with Convolutional Neural Networks. https://medium.com/swlh/computer-vision-with-convolutional-neural-networks

[4] Pawel Michalski,Bogdan Ruszczak,Michal Tomaszewski(2018). Convolutional Neural Networks Implementations for Computer Vision.

[5] Mikey Taylor (2020) Computer Vision with Convolutional Neural Networks. https://medium.com/swlh/computer-vision-with-convolutional-neural-networks

[5] Tairu Emmanual (2018).Plant Village Dataset https://www.kaggle.com/datasets/emmarex/plantdisease